MAXIMIZING THE LENGTH OF A SUCCESS RUN FOR MANY-ARMED BANDITS

by

Donald A. Berry*

and

Bert Fristedt**

University of Minnesota

Technical Report No. 379

May, 1981

*School of Statistics. Research supported by NSF Grant No. MCS80-01800.

**School of Mathematics. Research supported by NSF Grant No. MCS78-01168.

MAXIMIZING THE LENGTH OF A SUCCESS RUN FOR MANY-ARMED BANDITS

by

Donald A. Berry*

and

Bert Fristedt**

ABSTRACT

One of a number of Bernoulli processes is selected at each of a number of stages. A success at stage i is worth α_i and the problem is to maximize the expected payoff before the first failure. Results of Berry and Viscusi (1981) are generalized. In particular, we show that there is always an optimal strategy that uses a single process exclusively and indefinitely whenever the arms are independent and the discount sequence $(\alpha_1, \alpha_2, \ldots)$ is superregular. There is not always a similar reduction in the number of strategies when the discount sequence is not superregular.

Key Words: Many-armed bandits; Sequential decisions; Gambling with discounting; Bernoulli processes; Single-arm strategies; Stay-on-a-winner rule.

*School of Statistics (**School of Mathematics), University of Minnesota, Vincent Hall, 206 Church St. S.E., Minneapolis, MN 55455, U.S.A.

The first author's research was supported by NSF Grant No. MCS80-01800 and the second author's by NSF Grant No. MCS 78-01168.

1. Introduction

One of a number of Bernoulli processes (or "arms") is selected (or "pulled") at each of a (possibly infinite) number of stages. A success at stage i is worth α_i ($\alpha_i \geq 0$) and the problem is to maximize the expected payoff before the first failure. The discount sequence $(\alpha_1, \alpha_2, \ldots)$ is not necessarily monotone, as is usually assumed, nor is $\Sigma \alpha_i$ necessarily finite. Berry and Viscusi (1981) consider a number of discount sequences for which an optimal strategy is to use one arm exclusively and indefinitely (that is, one which "stays on a winner"). But they point out (their Example 4.2) that no such strategy is optimal for general discount sequences.

The present paper extends the Berry and Viscusi (1981) results. We show (Theorem 3.1) that there is an optimal strategy that uses a single arm whenever the arms are independent and the discount sequence is superregular (Definition 2.1). Superregularity is a more restrictive condition than that of regularity discussed by Berry and Fristedt (1979). The latter paper shows (in case the discount sequence in nonincreasing) that the stay-on-a-winner rule is optimal in the classical two-armed bandit with one arm known if and only if the discount sequence is regular. But in the present problem regularity is not enough, as is seen in Theorem 4.1. Even in the case of independence there are no neat characterizations of discount sequences for which the stay-on-a-winner rule is optimal: Example 4.1 shows that superregularity is not necessary, but Theorem 4.1 shows that it is almost necessary in a sense to be made clear later.

Applications of this problem are discussed by Berry and Viscusi (1981), Viscusi (1979a,1979b) and Viscusi and Zeckhauser (1976).

2. Preliminaries

Pulling arm j $(j = 1,\ldots,k)$ generates a Bernoulli random variable for which the probability of "success" is p_j. Given $p_1,\ldots,p_k$, such random variables are assumed to be mutually independent. The parameters $p_1,\ldots,p_k$ are themselves random variables with distribution measures $F_1,\ldots,F_k$, and in Sections 3 and 4 are assumed to be independent.

At each stage we are to pull one of the k arms. Let W_m be 1 if a success is obtained at stage m and 0 otherwise. For $m = 1,2,\ldots,$ define

$$Z_m = \prod_1^m W_i.$$

The probabilistic characteristics of the sequences $\{W_m\}$ and $\{Z_m\}$ depend on the strategy used. For a given discount sequence $\underset{\sim}{A} = (\alpha_1,\alpha_2,\ldots)$ we want to find a strategy that maximizes the expected <u>payoff</u>,

$$E \sum_1^\infty \alpha_m Z_m.$$

A strategy is called <u>optimal</u> if its expected payoff is

$$V = \sup E \sum_1^\infty \alpha_m Z_m,$$

where the supremum is over all possible strategies. An arm that is the first selection of an optimal strategy is also called optimal.

Because of the rather special nature of the problem, all sequential strategies of interest are very simple: since the only circumstance of interest occurs when all successes have been obtained, every strategy of interest is simply an infinite sequence of integers--the $m^{\underline{th}}$ integer indicating the arm to be selected at stage m.

We now show that there is an optimal strategy.

Lemma 2.1. Let

$$V_n = \sup E \sum_1^n \alpha_m Z_m,$$

where the supremum is over all possible strategies. Then $V_n \to V_\infty$ as $n \to \infty$.

Proof: Clearly, for any strategy,

$$\lim_{n\to\infty} E \sum_1^n \alpha_m Z_m \le \underline{\lim}_{n\to\infty} [\sup E \sum_1^n \alpha_m Z_m], \tag{2.1}$$

where the supremum is over all possible strategies. The result follows by taking the supremum on the left side of (2.1).□

Proposition 2.1. There exists an optimal strategy.

Proof: First suppose $V = \infty$. Clearly,

$$EZ_m = \prod_{j=1}^k Ep_j^{t_j}$$

where t_j is the number of times arm j is pulled during the first m stages. By the Hölder inequality and induction on k,

$$EZ_m \le \max_j Ep_j^m \le \sum_{j=1}^k Ep_j^m. \tag{2.2}$$

Therefore,

$$\infty = V = \sup \sum_{m=1}^\infty \alpha_m EZ_m \le \sum_{j=1}^k \sum_{m=1}^\infty \alpha_m Ep_j^m$$

and so for some j,

$$\sum_{m=1}^{\infty} \alpha_m Ep_j^m,$$

the expected payoff using arm j exclusively, equals $+\infty$. So, in fact, there is a single-arm strategy which is optimal.

Now suppose $V < \infty$. Since there are only finitely many strategies up to stage n, there exists an optimal strategy τ_n for $(\alpha_1,\ldots,\alpha_n,0,0,0,\ldots)$. We specify a strategy τ recursively. The first pull agrees with τ_n for infinitely many n. Make the first $m+1$ pulls of τ agree with the first $m+1$ pulls of τ_n for infinitely many n.

Let $\varepsilon > 0$. Since $V < \infty$,

$$\sum_{m=1}^{\infty} \alpha_m Ep_j^m < \infty$$

for each j. Hence, M can be chosen to make

$$\sum_{m=M}^{\infty} \alpha_m \sum_{j=1}^{k} Ep_j^m < \varepsilon. \tag{2.3}$$

In view of (2.2) and (2.3), if τ agrees with τ_n through stage M then τ is ε-optimal for $(\alpha_1,\ldots,\alpha_n,0,0,0,\ldots)$. Since τ does so agree with τ_n for infinitely many n and, by Lemma 2.1, $V_n \to V$, τ is ε-optimal for $\underset{\sim}{A}$. The conclusion follows since τ is ε-optimal for every $\varepsilon > 0$.□

For particular distributions, an optimal strategy is easy to specify. The following example is an immediate consequence of the first inequality in (2.2). It generalizes Theorem 3.1 of Berry and Viscusi (1981).

<u>Theorem 2.1</u>. Suppose $Ep_1^m \geq Ep_i^m$ for $i = 2,\ldots,k$ whenever $\alpha_m \neq 0$. Then it is optimal to pull arm 1 forever.

Corollary 2.1. If $F_1 = \ldots = F_k$ then all single-arm strategies are optimal.

The following concept is related to the notions of total positivity (Marshall and Olkin 1979, Ch. 18) and log-convexity.

Definition 2.1. A discount sequence $(\alpha_1, \alpha_2, \ldots)$ is superregular if, for all positive integers m and n,

$$\frac{\alpha_{m+1}}{\alpha_m} \geq \frac{\alpha_{m+n+1}}{\alpha_{m+n}},$$

which is understood to be satisfied whenever either side is $0/0$.

Remarks. All superregular discount sequences are unimodal. It follows that a discount sequence $\underset{\sim}{A} = (\alpha_1, \alpha_2, \ldots)$ is superregular iff both $\alpha_{m+1}^2 \geq \alpha_m \alpha_{m+2}$ for all m and all the members between any two positive members of $\underset{\sim}{A}$ are positive. The geometric sequence $(1, \alpha, \alpha^2, \ldots)$ is superregular for any $\alpha \geq 0$. Some other superregular sequences are $(0,0,1,1,\ldots,1,\alpha,\alpha^2,\ldots)$ for $0 \leq \alpha \leq 1$ and $(0,1,2,3,4,0,0,\ldots)$. Nonincreasing superregular sequences are "regular" as defined by Berry and Fristedt (1979), but regular sequences are not necessarily superregular; for instance, $(1,1,\frac{1}{2},\frac{1}{2},0,0,0,\ldots)$ is regular but not superregular.

In the next section we characterize optimal strategies when the discount sequence is superregular and the arms are independent.

3. Optimal Strategies for Superregular Sequences

The theorem in this section generalizes results of Berry and Viscusi (1981), who considered $k = 2$ and particular discount sequences. Examples 4.1 and 4.2 of Berry and Viscusi (1981), in which the discount sequences $(1,\alpha,\alpha^2,\ldots)$ and $(10,\alpha,\alpha^2,\ldots)$ for $\alpha \in (0,1)$ are considered, can be instructive for understanding the following theorem.

Theorem 3.1. Suppose $\underset{\sim}{A} = (\alpha_1,\alpha_2,\ldots)$ is superregular. Then, for any k independent arms, there is an optimal strategy that uses a single arm forever.

Remark. To see that the assumption of independence is necessary, suppose $\underset{\sim}{A} = (1,\alpha,\alpha^2,\ldots)$, $\alpha \in (0,1)$, $k = 2$ and (p_1,p_2) is known to be either $(3/4,1)$ or $(1/4,0)$. If the probability of $(3/4,1)$ is less than $(1-\alpha)/(2-\alpha)$ then an optimal strategy must begin with a pull of arm 1 and must use arm 2 when the probability of $(3/4,1)$ becomes greater than $(1-\alpha)/(2-\alpha)$.

For the proof of Theorem 3.1 we shall use the following result.

Lemma 3.1. For any $c \in [0,1]$ and any random variable X with distribution measure F, there exists an $M = M(F,c) \in [1,\infty]$ such that, for positive integers m,

$$EX^{m-1}(X-c)\begin{cases} < 0 & \text{if } m < M \\ \geq 0 & \text{if } m \geq M \end{cases} .$$

Proof: For $u > 1$ define

$$g_c(u) = EX^{u-1}(X-c),$$

which is an analytic function for any $c \in [0,1]$. Then

$$g_c'(u) = EX^{u-1}(X - c)\log X$$
$$= EX^{u-1}(X - c)(\log X - \log c) + g_c(u)\log c.$$

If $g_c(u) \leq 0$ then $g_c'(u) \geq 0$ since $\log c \leq 0$ and

$$x^{u-1}(x - c)(\log x - \log c) \geq 0$$

for all $x \in [0,1]$ and $c \in [0,1]$. The result follows from the continuity of g_c on $(1,\infty)$ and the fact that $g_c(1+) \geq g_c(1)$.□

<u>Proof of Theorem 3.1</u>: Let S_n denote the set of all superregular discount sequences $(\alpha_1,\alpha_2,\ldots)$ satisfying the conditions $\alpha_n > 0$ and $\alpha_{n+1} = \alpha_{n+2} = \ldots = 0$. We proceed by induction on n.

If $\underset{\sim}{A} \in S_1$ the result is trivial. For $n \geq 2$ assume that the result holds for every member of S_{n-1}. We will show that it holds for every member of S_n.

If $\underset{\sim}{A} = (\alpha_1,\alpha_2,\ldots) \in S_n$ then $(\alpha_2,\alpha_3,\ldots) \in S_{n-1}$. By Proposition 2.1 and the inductive hypothesis, there is an optimal strategy τ that uses a single arm at stages 2 through n. So τ uses at most two arms. If it uses only one arm the conclusion follows. Suppose it uses two arms, say arm 1 initially and arm 2 thereafter.

Since τ is optimal it is at least as good as τ_1: "Pull arm 1 at stages 1 and 2 and arm 2 thereafter." Thus, the difference in expected payoffs of τ and τ_1 is nonnegative; that is,

$$\alpha_1 Ep_1 + Ep_1 \sum_{m=2}^{\infty} \alpha_m Ep_2^{m-1} - \alpha_1 Ep_1 - Ep_1^2 \sum_{m=2}^{\infty} \alpha_m Ep_2^{m-2} \geq 0.$$

So using $Ep_1^2 \geq (Ep_1)^2$,

(3.1) $$\sum_{m=1}^{\infty} \alpha_{m+1} Ep_2^{m-1}(p_2 - Ep_1) \geq 0.$$

Let $M = M(F_2, Ep_1)$ as defined in Lemma 3.1. From (3.1), $n \geq M + 1$; hence, since $\alpha_n > 0$, we may define

$$N = \min\{m : m \geq M,\ \alpha_{m+1} > 0\}.$$

In view of (3.1), we have

(3.2) $$\sum_{m=1}^{\infty} \alpha_N \alpha_{m+1} Ep_2^{m-1}(p_2 - Ep_1) \geq 0.$$

By superregularity,

$$\alpha_{N+1}\alpha_m \leq \alpha_N \alpha_{m+1} \quad \text{for} \quad m < M$$

$$\alpha_{N+1}\alpha_m \geq \alpha_N \alpha_{m+1} \quad \text{for} \quad m \geq M.$$

Therefore, (3.2) implies, using Lemma 3.1,

(3.3) $$\sum_{m=1}^{\infty} \alpha_{N+1}\alpha_m Ep_2^{m-1}(p_2 - Ep_1) \geq 0.$$

So,

$$Ep_1 \sum_{m=1}^{\infty} \alpha_m Ep_2^{m-1} \leq \sum_{m=1}^{\infty} \alpha_m Ep_2^m .$$

That is, the expected payoff using τ is no larger than that of pulling arm 2 exclusively, a strategy that is therefore optimal.

It remains to consider the case in which $\alpha_m > 0$ for infinitely many m. The proof of Proposition 2.1 applies easily where the τ_n given there are chosen to be single-arm strategies.□

4. Non-Superregular Discount Sequences

As in the previous section we assume that the arms are independent. When $\underset{\sim}{A}$ is not superregular then without knowing $(F_1,\ldots,F_k)$ it may or may not be possible to say that there is a single-arm strategy that is optimal. If the condition of superregularity fails in the first three stages then Theorem 4.1 says that there are distributions F_j for which there is no such optimal strategy. However, if it does not fail in the first three stages but does later, then, as Example 4.1 shows, there may always be an optimal single-arm strategy.

Theorem 4.1. For any k independent arms, if $\underset{\sim}{A}$ is such that $\alpha_1\alpha_3 > \alpha_2^2$, then there are distributions $F_1,\ldots,F_k$ for which all optimal strategies use at least two arms.

Proof: Without loss we can assume $k = 2$. We shall show that the class of distributions indexed by x and r as follows:

$$F_1(\{x\}) = 1$$

$$F_2(\{2x\}) = r, \quad F_2(\{0\}) = 1 - r,$$

contains a member which satisfies the theorem.

Consider three strategies, σ_1: "Pull arm 1 exclusively," σ_2: "Pull arm 2 exclusively," and σ: "Pull arm 1 initially and arm 2 thereafter." Let D_i, $i = 1,2$, equal the expected payoff using σ minus the expected payoff using σ_i. For $\delta > 0$, $x \downarrow 0$ and

$$r = \tfrac{1}{2} - \frac{\alpha_3 x}{2(\alpha_2 + \delta)} \quad ,$$

$$D_1 = \alpha_2(2r - 1)x^2 + \alpha_3(4r - 1)x^3 + O(x^4) = \frac{\delta\alpha_3}{\alpha_2 + \delta}x^3 + O(x^4)$$

and

$$D_2 = \alpha_1(1 - 2r)x - 2\alpha_2 rx^2 + 0(x^3)$$

$$= \frac{\alpha_1\alpha_3 - \alpha_2^2 - \delta\alpha_2}{\alpha_2 + \delta} x^2 + 0(x^3) \ .$$

Since both D_1 and D_2 are positive for an appropriate fixed δ and x sufficiently small, neither σ_1 nor σ_2 is optimal.□

In Theorem 4.1 superregularity was assumed to fail in the first three stages. If, instead, it fails in stages 2, 3, and 4, one might hope to prove a similar result by arranging for the right kind of distribution on $(p_1,\ldots,p_k)$ at stage 2. However, that turns out to be impossible; indeed, as the following example shows, all optimal strategies may be one-arm strategies in the absence of superregularity.

Example 4.1. Consider the discount sequence $(1,1,\varepsilon,\varepsilon,0,0,0,\ldots)$ which is not superregular unless $\varepsilon = 0$ or 1 (and not regular (Berry and Fristedt 1979) for $0 < \varepsilon < .5$). We shall show, for $0 < \varepsilon \leq .5$, that there is always a one-arm strategy that is optimal.

Since we are still assuming independence we conclude from Theorem 3.1 that there is an optimal strategy that indicates the same arm after stage 3 as at stage 3. Accordingly, we need consider no more than three arms -- say arms 1, 2, and 3. We use hij where each of h, i, j equals 1, 2, or 3 to denote the strategy: "Pull arm h at stage 1, arm i at stage 2, and arm j thereafter," and V(hij) to denote the expected payoff for the strategy hij. We want to prove that V takes on its maximum at one of the strategies 111, 222, 333.

In case h, i, j are all distinct either $V(hij) \leq V(hhj)$ or $V(hij) \leq V(iij)$ by application of Theorem 3.1 for the discount sequences $(1, 1+\varepsilon Ep_j + \varepsilon Ep_j^2, 0, 0, 0, \ldots)$. Thus, we may, and do, restrict out attention to strategies involving at most two arms -- say, arm 1 and arm 2, and we assume $Ep_1 \geq Ep_2$.

There are then six strategies which we claim are no better than one-arm strategies, a claim which will follow from:

(i) $V(121) \leq V(111)$,

(ii) $V(211) \leq V(121)$,

(iii) $V(112) \leq V(111) \vee V(222)$,

(iv) $V(221) \leq V(111) \vee V(222)$,

(v) $V(122) \leq V(112) \vee V(222)$,

(vi) $V(212) \leq V(122)$.

We will demonstrate (iii), (iv), and (v) and leave the easier (i), (ii), and (vi) for the reader.

Let μ_m and ν_m denote the $m^{\underline{th}}$ moments of F_1 and F_2.

The following calculation proves (iii) in case $\nu_1 + \nu_2 \leq \mu_1 + \mu_2$:

$$
\begin{aligned}
V(112) &= \mu_1 + \mu_2 + \varepsilon\mu_2\nu_1 + \varepsilon\mu_2\nu_2 \\
&\leq \mu_1 + \mu_2 + \varepsilon\mu_2(\mu_1 + \mu_2) \\
&\leq \mu_1 + \mu_2 + \varepsilon(\mu_3 + \mu_4) = V(111),
\end{aligned}
$$

the inequalities $\mu_2\mu_1 \leq \mu_3$ and $\mu_2^2 \leq \mu_4$ following from the Cauchy-Schwarz inequality. If $\nu_1 + \nu_2 > \mu_1 + \mu_2$ then $\nu_2 > \mu_2$ (since $\nu_1 \leq \mu_1$) and

$$
\begin{aligned}
V(112) &< \nu_1 + \nu_2 + \varepsilon\nu_2(\nu_1 + \nu_2) \\
&\leq \nu_1 + \nu_2 + \varepsilon(\nu_3 + \nu_4) = V(222).
\end{aligned}
$$

The following calculation proves (iv) in case $\nu_2 \le \mu_2$:

$$\begin{aligned} V(221) &= \nu_1 + \nu_2 + \varepsilon\nu_2\mu_1 + \varepsilon\nu_2\mu_2 \\ &\le \mu_1 + \mu_2 + \varepsilon\mu_2(\mu_1 + \mu_2) \\ &\le \mu_1 + \mu_2 + \varepsilon(\mu_3 + \mu_4) = V(111). \end{aligned}$$

In case $\nu_2 > \mu_2$, using the Cauchy-Schwarz inequality,

$$\begin{aligned} V(222) - V(221) &= \varepsilon(\nu_3 - \nu_2\mu_1 + \nu_4 - \nu_2\mu_2) \\ &\ge \varepsilon\nu_2(\frac{\nu_2}{\nu_1} - \mu_1 + \nu_2 - \mu_2) \\ &\ge \varepsilon\nu_2(\frac{\mu_2}{\mu_1} - \mu_1) \ge 0. \end{aligned}$$

We now show (v). First assume $\mu_1 \ge \nu_2/\nu_1$; we will show $V(122) \le V(112)$. To do this we calculate

$$\begin{aligned} \frac{\nu_1}{\mu_1}[V(112) - V(122)] &= \frac{\nu_1}{\mu_1}[\mu_2 - \mu_1\nu_1 + \varepsilon(\mu_2\nu_1 - \mu_1\nu_2 + \mu_2\nu_2 - \mu_1\nu_3)] \\ &\ge \nu_1\mu_1 - \nu_1^2 + \varepsilon\nu_1(\mu_1\nu_1 - \nu_2 + \mu_1\nu_2 - \nu_3), \end{aligned} \tag{4.1}$$

which is an increasing function of μ_1, and therefore we need only consider $\mu_1 = \nu_2/\nu_1$. In this case (4.1) becomes

$$\begin{aligned} \nu_2 - \nu_1^2 + \varepsilon(\nu_2^2 - \nu_1\nu_3) &\ge \nu_2 - \nu_1^2 + \varepsilon(\nu_1^2\nu_2 - \nu_1\nu_3) \\ &\ge \nu_2 - \nu_1^2 + \varepsilon(\nu_1\nu_2 - \nu_3) = (\nu_2 - \varepsilon\nu_3) - \nu_1(\nu_1 - \varepsilon\nu_2) \end{aligned}$$

which, being the covariance of p_2 and an increasing function, $p_2 - \varepsilon p_2^2$, of p_2, is nonnegative (Lehmann 1966). It remains to consider the case

$\mu_1 < \nu_2/\nu_1$, which by the Cauchy-Schwarz inequality implies

$$(4.2) \qquad \mu_1 < \nu_3/\nu_2 \leq \nu_4/\nu_3 .$$

We calculate

$$(4.3) \qquad V(222) - V(122) = \nu_1 - \mu_1 + \nu_2 - \mu_1\nu_1 + \varepsilon(\nu_3 - \mu_1\nu_2 + \nu_4 - \mu_1\nu_3).$$

In view of (4.1), (4.2), and (4.3),

$$\frac{1}{\mu}[V(112) - V(122)] + [V(222) - V(122)]$$
$$\geq (1 - \varepsilon)(\nu_2 - \mu_1\nu_1) + \varepsilon(\nu_4 - \mu_1\nu_3),$$

which by (4.2) is positive. Therefore, either $V(112) \geq V(122)$ or $V(222) \geq V(122)$.□

References

Berry, D. A. and Fristedt, B. (1979). Bernoulli one-armed bandits--arbitrary discount sequences. Ann. Statist. 7, 1086-1105.

Berry, D. A. and Viscusi, W. K. (1981). Bernoulli two-armed bandits with geometric termination. Stochastic Proc. and Their Applic. 11, 35-45.

Lehmann, E. L. (1966). Some concepts of dependence. Ann. Math. Statist. 37, 1137-1153.

Marshall, A. W. and Olkin, I. (1979). Inequalities: Theory of Majorization and Its Applications. Academic Press, New York.

Viscusi, W. K. (1979a). Job hazards and worker quit rates: An analysis of adaptive worker behavior. International Econ. Rev. 20, 29-58.

Viscusi, W. K. (1979b). Employment Hazards: An Investigation of Market Performance. Harvard University Press, Cambridge.

Viscusi, W. K. and Zeckhauser, R. (1976). Environmental policy change under uncertainty. J. Environmental Economics and Management 3, 97-112.